The Credibility Crisis and Computational Science: Accountability and Public Health

Victoria Stodden
Department of Statistics
Columbia University

NYSPI Biostatistics Seminar Series
Columbia University
New York
September 20, 2011

Computational Methods Emerging as Central to the Scientific Enterprise

1. enormous, and increasing, amounts of data collection,
   - ~3TB/yr genome sequence data: ~1000 sequencers running full time producing 600GB each run (HiSeq 2000, 11 days per run),
   - CMS project at LHC: 300 "events" per second, 5.2M seconds of runtime per year, .5MB per event = 780TB/yr => several PB when data processed,
   - Sloan Digital Sky Survey: 8th data release (2010), 49.5TB.
2. massive simulations of the complete evolution of a physical system, systematically varying parameters,
3. deep intellectual contributions now encoded in software.

Updating the Scientific Method

Donoho and others have argued that computation presents only a *potential* third branch of the scientific method:

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments,
- Branch 3? (computational): large scale simulations.

The Ubiquity of Error

- The central motivation for the scientific method is to root out error:
  - Deductive branch: the well-defined concept of the proof,
  - Empirical branch: the machinery of hypothesis testing, structured communication of methods and protocols.
- Computational science as practiced today does not generate reliable knowledge.
- See e.g. Ioannidis, "Why Most Published Research Findings are False," PLoS Med, 2005.

Computation Emerging as Central to the Scientific Endeavor

JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

- Data and code typically not made available at the time of scientific publication, rendering results unverifiable, not reproducible.

➡ A *Credibility Crisis*

Duke Clinical Trials Case

- Potti et al (2006), Nature Medicine; (2006) NEJM; (2007) Lancet Oncology; (2007) Journal of Clinical Oncology: evidence of genomic signatures to guide use of chemotheraputics (*all since retracted*),
- Coombes, Wang, Baggerly at M.D. Anderson Cancer Center cannot replicate, and find simple flaws: genes misaligned by one row, column labels flipped, genes repeated and missing from analysis..

Correcting the Flaws

- 2007 correspondence and a supplementary report submitted to the Journal of Clinical Oncology - publication declined; 2008 Nature Medicine declines to publish correspondence.
- Clinical trials initiated in 2007 (Duke), 2008 (Moffitt).
- Duke launches internal investigation Sept 2009; all three trials suspended in Oct 2009,
- Oct 2009: results reported validated, regardless of errors, because data blinded (later found not to be true),

Clinical Trials?

- Jan 2010: Duke clinical trials resume, patients allocated to treatment and control groups. "Neither the review nor the raw data are being made available at this time."
- July 2010: 33 prominent biostatisticians write to Harold Varmus as NCI Director urging suspension of the trials and an examination of standards of review, including reproducibility.
- Sept 2010: IOM committee "Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials" formed,
- Nov 2010: Potti resigns and the clinical trials are terminated.

Excerpt: Letter to Varmus

"We strongly urge that the clinical trials in question ... be suspended until a fully independent review is conducted of both the clinical trials and of the evidence and predictive models being used to make cancer treatment decisions. For this to happen, sufficiently detailed data and annotation must be made available for review. The data should be sufficiently documented for provenance to be assessed (as both gene and sample mislabeling have been documented in these data), and the computer code used to predict which drugs are suitable for particular patients must be made available to allow an independent group of expert genomic data analysts to assess its validity and reproducibility using the data supplied."

INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

ABOUT THE IOM | REPORTS | ACTIVITIES

Browse History

Activity

Bookmark this Page | Print | E-mail | ShareThis

Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials

Type: Consensus Study
Topics: Biomedical and Health Research, Health Services, Coverage, and Access
Boards: Board on Health Care Services

Activity Description

An IOM committee will review the published literature to identify appropriate evaluation criteria for tests based on "omics" technologies (e.g. genomics, epigenomics, proteomics, metabolomics) that are used as predictors of clinical outcomes. The committee will recommend an evaluation process for determining when predictive tests based on omics technologies are fit for use as a basis for clinical trial design, including stratification of patients and response to therapy in clinical trials. The committee will identify criteria important for the analytical validation, qualification, and utilization components of test evaluation.

The committee will apply these evaluation criteria to predictive tests used in three cancer clinical trials conducted by Duke University investigators (NCT00509366, NCT00545948, NCT00636441). For example

Duke Recent Events

- IOM Committtee meetings, including with Duke representatives on August 22, 2011,
- 2 lawsuits filed by patients in the clinical trials, asserting that "that Duke's response "to the accusation of invalid and fraudulent science was deceptive, misleading, and fraudulent conduct designed to protect its reputation and proprietary interests … rather than protecting the safety of the patients involved in the clinical trials." This "reduced the Plaintiffs' likelihood of surviving his/her cancer or the likelihood of experiencing a positive response to the chemotherapy regimen."

Framing Principle for Scientific Communication: *Reproducibility*

- "The idea is: An article about computational science in a scientific publication is *not* the scholarship itself, it is merely *advertising* of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures." David Donoho, 1998.

- (simple) definition: a result is reproducible if a member of the field can independently verify the result.

Central Thesis

- *Computational science must develop standards for reproducibility before it can be considered a third branch of the scientific method,*

➡Data and Code sharing with publication.

The Role of Policy

Congress: Bayh-Dole Act

- Bayh-Dole Act (1980), designed to promote the transfer of academic discoveries for commercial development, via licensing of patents.
- Legislators blind to the coming digital revolution, impact on software and algorithm patenting. Tech Transfer Offices and code release.
- Implications for science as a disruptor of openness norms:
    - patents => delay in revealing code, or closed code,
    - I assert Bilski => obfuscation of methods submitted for patents,
    - (aside from altering a scientist's incentives toward commercial ends).

Congress: America COMPETES

- America COMPETES Re-authorization (2011):
  - § 103: Interagency Public Access Committee:

    "coordinate Federal science agency research and policies related to the dissemination and long-term stewardship of the results of unclassified research, *including digital data* and peer-reviewed scholarly publications, supported wholly, or in part, by funding from the Federal science agencies." (emphasis added)

  - § 104: Federal Scientific Collections: OSTP "shall develop policies for the management and use of Federal scientific collections to improve the quality, organization, *access, including online access*, and long-term preservation of such collections for the benefit of the scientific enterprise." (emphasis added)

Funding Agency Policy

- NSF grant guidelines:

  "NSF ... expects investigators to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work. It also encourages grantees to share software and inventions or otherwise act to make the innovations they embody widely useful and usable."

- NSF peer-reviewed Data Management Plan, January 2011.
- NIH (2003): "The NIH endorses the sharing of final research data to serve these and other important scientific goals. The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers." (>$500,000, include data sharing plan)

Journal Policy

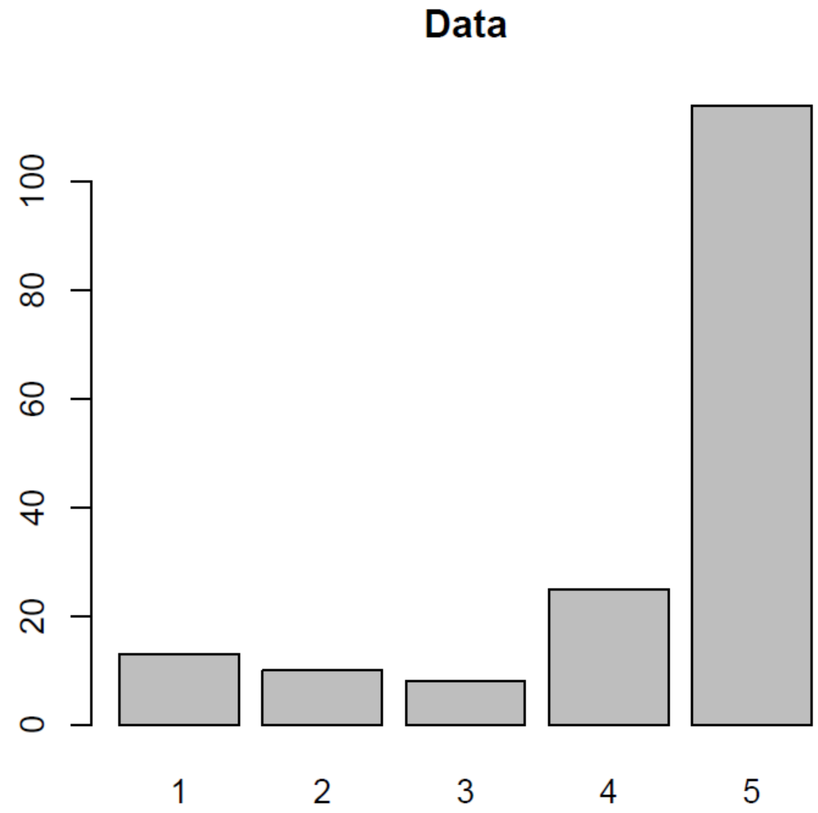
- Different approaches by journals:
  - may offer unreviewed "supplemental materials" section,
  - may require data and/or code to be provided upon request (Science as of Feb 11 2011),
  - may employ an Associate Editor for Reproducibility (Biostatistics, Biometrical Journal) or replicate results (ACM SIGMOD),
  - may publish correspondence from the review process (Molecular Systems Biology, The European Molecular Biology Organization Journal),
  - new journals, ie. Open Research Computation, BMC Data Notes
  - ignore the issue entirely..

Journal Policy: Preliminary Results

- ISI Journal Citation Report
  - Impact Factor (citations/articles)
- Fields:
  - Mathematical & Computational Biology
  - Statistics & Probability
  - Multidisciplinary Science
- 176 journals examined, 170 studied

Data Policy

Data		
Mark	Count	Percentage
1	13	0.0765
2	10	0.0588
3	8	0.0471
4	25	0.1471
5	114	0.6706
Total	170	1.0000



Data Sharing Mark
1 - required, affects publication decisions or accesion number required
2 - required, but no mention of effect on publication, accession number may not be strictly required
3 - not required, both reviewed and hosted
4 - not required, either not reviewed or not hosted or no mention of either
5 - no explicit

with Peixuan Guo

Groundswell from across the Computational Sciences

- AMP 2011 "Reproducible Research: Tools and Strategies for Scientific Computing"
- AMP / ICIAM 2011 "Community Forum on Reproducible Research Policies"
- SIAM Geosciences 2011 "Reproducible and Open Source Software in the Geosciences"
- ENAR International Biometric Society 2011: Panel on Reproducible Research
- AAAS 2011: "The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer"
- SIAM CSE 2011: "Verifiable, Reproducible Computational Science"
- Yale 2009: Roundtable on Data and Code Sharing in the Computational Sciences
- ACM SIGMOD conferences
- NSF/OCI report on Grand Challenge Communities (Dec, 2010)
- IOM "Review of Omics-based Tests for Predicting Patient Outcomes in Clinical Trials"
- ...

Final Observations

- Reproducibility emerging as a movement, with corollary efforts to enable code and data sharing.
  - interlocking incentives: federal policy, funding agency policy, journal policy, review and tenure committees.
  - facilitation of reproducibility through tool development (see http://www.stodden.net/AMP2011).
  - serious consequences, both in the short term (e.g. clinical trials) and in the long term (establishing scientific facts).

References

- “Enabling Reproducible Research: Open Licensing for Scientific Innovation”
- “The Scientific Method in Practice: Reproducibility in the Computational Sciences”
- “Open Science: Policy Implications for the Evolving Phenomenon of User-led Scientific Innovation”
- Reproducible Research: Tools and Strategies for Scientific Computing, July 2011
- Reproducible Research in Computational Science: What, Why and How, Community Forum, July 2011

available at http://www.stanford.edu/~vcs

Supplemental Slides

Legal Barriers: Copyright

"To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." (U.S. Const. art. I, §8, cl. 8)

- Original expression of ideas falls under copyright *by default* (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original
  - limited time: generally life of the author +70 years

Exceptions and Limitations: Fair Use.

Responses Outside the Sciences 1: Open Source Software

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

- Hundreds of open source software licenses:
  - GNU Public License (GPL)
  - (Modified) BSD License
  - MIT License
  - Apache 2.0 License
  - ... see http://www.opensource.org/licenses/alphabetical

Responses Outside the Sciences 2: Creative Commons

- Founded in 2001, by Stanford Law Professor Larry Lessig, MIT EECS Professor Hal Abelson, and advocate Eric Eldred.
- Adapts the Open Source Software approach to artistic and creative digital works.

Responses Outside the Sciences 2: Creative Commons

- Creative Commons provides a suite of licensing options for digital artistic works:
  - BY: if you use the work attribution must be provided,
  - NC: the work cannot be used for commercial purposes,
  - ND: no derivative works permitted,
  - SA: derivative works must carry the same license as the original

Response from Within the Sciences

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

- A suite of license recommendations for computational science:
  - Release media components (text, figures) under CC BY,
  - Release code components under Modified BSD or similar,
  - Release data to public domain or attach attribution license.

➡ Remove copyright's barrier to reproducible research and,

➡ Realign the IP framework with longstanding scientific norms.

Winner of the Access to Knowledge Kaltura Award 2008

Advantages to the RRS

- focus becomes the release of the entire research compendium,
- hook for funding agencies, journals, universities,
- standardization avoids license incompatibilities,
- clarity of rights (beyond fair use),
- this IP frameworks supports scientific norms.

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%